A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy

Michael J. Sanderson
Section of Evolution and Ecology, University of California, Davis

A new method for estimating divergence times when evolutionary rates are variable across lineages is proposed. The method, called nonparametric rate smoothing (NPRS), relies on minimization of ancestor–descendant local rate changes and is motivated by the likelihood that evolutionary rates are autocorrelated in time. Fossil information pertaining to minimum and/or maximum ages of nodes in a phylogeny is incorporated into the algorithms by constrained optimization techniques. The accuracy of NPRS was examined by comparison to a clock-based maximum-likelihood method in computer simulations. NPRS provides more accurate estimates of divergence times when (1) sequence lengths are sufficiently long, (2) rates are truly nonclocklike, and (3) rates are moderately to highly autocorrelated in time. The algorithms were applied to estimate divergence times in seed plants based on data from the chloroplast *rbcL* gene. Both constrained and unconstrained NPRS methods tended to produce divergence time estimates more consistent with paleobotanical evidence than did clock-based estimates.

Introduction

Since the protein “molecular clock” was first postulated (Zuckerkandl and Pauling 1962, 1965) and the neutral theory was proposed (Kimura 1983), numerous attempts to reconstruct the ages of clades from sequences evolving at roughly constant rates have been undertaken. Now that extensive evidence on molecular evolutionary rates has accumulated, much of which demonstrates rate variation across lineages rather than constancy (Britten 1986; Avise 1994; Li 1997), it is perhaps surprising that interest in estimating ages has not diminished. Indeed, if anything, it has increased in the past few years. Major recent studies have estimated the ages of eukaryotes (Doolittle et al. 1996), metazoans (Wray, Levinton, and Shapiro 1996), mammals and birds (Hedges et al. 1996; Cooper and Penny 1997), angiosperms (Goremykin, Hansmann, and Martin 1997), and basidiomycetes (Berbee and Taylor 1993) as well as many smaller clades (e.g., Russo, Takezaki, and Nei 1995; Kooistra and Medlin 1996). Unlike some earlier works, many of these studies tested rate constancy prior to the use of particular genes to estimate divergence times. However, studies using the most rigorous tests of rate constancy have tended to reject a large subset of possible genes because of departures from a clock (e.g., Hedges et al. 1996; Gu 1997), reducing the amount of data that can be brought to bear on the problem. Large data sets are almost essential for narrowing confidence intervals associated with the Poisson processes used to model sequence evolution (Cox and Miller 1977; Rodrigo et al. 1990). Although the variance of divergence time estimates can also be reduced by pooling pairwise comparisons from multiple taxa within clades (Steel, Cooper, and Penny 1996), this assumes that all of the pooled lineages are clocklike. Other variance-reduction techniques are likely to suffer from the same problem. Unfortunately, it is likely that even if a number of rate constant lineages are found, addition of more to an analysis will eventually uncover ones that deviate locally from a clock because of unique events in the history of that lineage. Thus, it is desirable to develop and test methods that do not assume global rate constancy.

Existing Methods for Estimation of Divergence Times in the Absence of Rate Constancy

Existing methods for reconstructing divergence times in the absence of a molecular clock fall into two broad categories. In the first, lineages are selectively removed if they deviate from a “backbone” that is clocklike. For example, Li and Tanimura (1987) selectively pruned away lineages that deviated from rate constancy in a series of relative-rate tests in primates, and then reconstructed divergence times using remaining lineages under the assumption that they were clocklike. Takezaki, Rzhetsky, and Nei (1995) formalized this procedure by using a nested series of relative-rate tests and combining information from multiple sequences within clades to eliminate lineages that deviated significantly from rate constancy. They also used the deviations of the total branch length from root to tip from the mean length to prune lineages. They referred to these pruned trees as “linearized” trees, not because they look untreelike, but rather because they make the remaining divergence times linear with distance. Russo, Takezaki, and Nei (1995) used this method to reconstruct divergence times in Hawaiian Drosophilids; Hedges et al. (1996) used it to reconstruct the timing of the diversification of orders of mammals; and Kooistra and Medlin (1996) used it to reconstruct the age of diatoms. Since it relies on relative-rate tests, the linearized tree approach may be sensitive to known problems with such tests, especially inability to detect parallel rate changes (Fitch 1976); however, the significance of this problem in practice is unknown.

The second approach incorporates rate variation directly into inference procedures. Hasegawa, Kishino, and Yano (1989) and Kishino and Hasegawa (1990) developed a distance-based approach that allowed different
rate parameters in different parts of the tree. In a four-taxon tree of primates, they allowed three pairs of transition/transversion rates in different lineages. Maximum likelihood was used to simultaneously estimate the three different rate pairs and associated divergence times.

Uyenoyama (1995) studied rates and divergence times in phylogenies of self-incompatibility alleles in plants using a similar framework. She argued that, especially in the case of distinct functional classes of sequences, it was reasonable to assume that functional constraints might alter rates in a lineage-specific manner. Her model assigned different lineage rates to different functional classes of alleles and demonstrated that the age of sporophytic self-incompatibility alleles in *Brassica* is several times older than the genus itself. As with Kishino and Hasegawa’s (1990) work, the supposition was that there might be “locally” constant rates in part of a tree despite rate variation at a larger phylogenetic scale.

These are so-called “model selection” methods, which implicitly or explicitly force the investigator to choose from among a possibly large set of potentially relevant models. In phylogenetics the difficult issue is how to set up an appropriate model with rate parameters assigned to different branches or sets of branches. For example, in the case of the four-taxon, three-parameter model used to study primate rates (Hasegawa, Kishino, and Yano 1989), three parameters can be assigned among seven branches. There are 301 ways to construct three-parameter models in a tree of seven branches (given by Stirling numbers of the second kind; Bogart 1990). Worse yet, the number of all possible models is much larger, because models might have not just three, but from 1 to *n* parameters, where *n* is the number of branches. The number of all possible models is equal to the number of partitions of a set of *n* objects (Bogart 1990), which for seven branches is 877 different models. Obviously, this approach has to be guided by strict prior assumptions about where rates are likely to change on the tree, so that the number of distinct models can be limited.

There are also mathematical pathologies associated with some models. It is not always possible to estimate the rates and times for some internal branches that have unique rates, because any rate increase could be compensated by a decrease in branch duration to keep the overall likelihood of the model the same. This is the often discussed degeneracy associated with the fact that sequence differences emerge as a function of the product of rate and time (Felsenstein 1981). Such problems can be avoided by selecting models in which rates are associated with branches in such a way that they end in tips or the root or are contiguous with branches that end in tips or the root and hence are “anchored” firmly in fixed time points (Uyenoyama 1995). This is true of the rate parameters in the primate case described by Hasegawa, Kishino, and Yano (1989). Without this anchoring, the likelihood function would certainly tend to be flat in some directions, leading to degenerate parameter estimates.

A priori, there is often no biological or evolutionary reason to expect the rate of evolution to be different in one subregion(s) of a tree versus another. Consequently, it will often be difficult to justify the study of any particular model or manageable set of models from among the many models possible for a given tree. Moreover, the degeneracies entailed by certain models would have to be removed by some method, or the models would have to be excluded from consideration. Alternatively, it is possible to build into the estimation procedure some constraint that will allow the data themselves to suggest when shifts in evolutionary rate are most likely to have occurred.

**Autocorrelation in Rate**

One possible approach is to place a constraint on the way rates can vary across a clade. Current practice, embodied in most recent empirical studies of divergence times, already imposes a very strong constraint by assuming a constant rate across a tree. This paper replaces that constraint with a much weaker constraint on how rates vary, but one that is still sufficient to allow estimation of divergence times. The constraint is the temporal autocorrelation of rates in lineages. An autocorrelation puts limits on the speed with which a rate can change from an ancestral lineage to a descendant lineage.

The notion of a temporal autocorrelation in rates of molecular evolution is not a new one. Gillespie (1991, p. 140), for example, in discussing the empirical evidence for a progressive slowdown in rates of evolution in primates, suggested that “. . . the rate of substitution evolves along lineages. Daughter lineages might inherit their initial rate from their parental lineage and evolve new rates independently.” He went on to argue that this might occur, for example, if something like Takahata’s “fluctuating neutral space” model of molecular evolution held, in which substitutions in a sequence alter the rate of evolution of the molecule itself (Gillespie 1991, pp. 141, 280; Takahata 1987). However, this may not be necessary, as any kind of temporal autocorrelation would suffice, including that induced by extrinsic environmental autocorrelation in selection intensities or other population level processes (Chao and Carr 1993).

**Methods**

The general problem is to estimate a set of divergence times in a phylogenetic tree in which the rate of evolution may vary between lineages. Such “lineage”-specific effects should be distinguished from locus-specific effects or variation in rate among sites within a gene at a single locus, neither of which is considered here. Unless some part of the tree is calibrated with an absolute date (using a fossil or geologic event), estimated times are relative to the unknown root node time of the tree.

**Nonparametric Rate Smoothing**

The method proposed in this paper will first be stated in a very general form. Then the remainder of this paper will focus on a simple specialization of this meth-
A Simple Estimator of Local Rate

The simplest local estimate of rate is just

\[ \hat{\lambda}_k = b_k / T_k, \]

where \( b_k \) is the length of branch \( k \) and \( T_k \) is its temporal duration. This local estimate of rate can be justified on a number of grounds. First, it is the standard definition of an observational rate as a change in a quantity during an interval of time, \( \Delta x / \Delta t \). Second, it is the maximum-likelihood estimator of rate in a Poisson process. In a Poisson process, the number of substitutions in an interval \( T_k \) follows a Poisson distribution, and \( \hat{\lambda}_k \) is a maximum-likelihood (and unbiased) estimator of the rate parameter (Evans, Hastings, and Peacock 1993).

Estimation of local rate based on such a small “window” (one branch) may be sensitive to noise in short branches. In principle, the window can be expanded such that an estimate of local rate for a branch might be determined by neighboring branches beyond the immediate descendants. This is similar to procedures in nonparametric regression in which a sliding window for estimating local regression slope can be scaled depending on the noisiness of the data (Härdle 1990).

Transformation in Rate at the Root Node of the Tree

As in most estimation problems associated with phylogenetic inference, there is a complication associated with estimation around the root of the tree. In likelihood tree inference procedures, for example, the root node frequency of nucleotide bases is usually estimated heuristically by the mean frequency over extant taxa (Felsenstein 1981; Swofford et al. 1996). In the formulation above, the transformation in rate from the branch below the root node (the “root” branch) to the two branches descended from the root (its “descendant” branches) is ignored because we do not have any information about the local rate of the root branch. In effect, this breaks the optimization problem into two separate problems for the two descendant subclades. It is possible that one subclade might have a very much higher rate estimate than the other, which is unlikely if there is any autocorrelation of rate from the root branch to the two descendant branches. One solution would be to clamp the difference in the two descendant rate estimates (which can be calculated) by a term like that in equation...
(1) above. Another solution, which is adopted here, is to estimate the root branch local rate as the mean of all the estimated descendant rates throughout the tree and then construct an expression analogous to equation (1) for the root node. Let the mean estimated rate be

\[ \hat{r}_{\text{root}} = \frac{1}{n} \sum_k \hat{r}_k, \]  

(4)

where \( n \) is the number of branches, and then the objective function, \( W \), must have the following term added to it.

\[ w_{\text{root}} = \sum_{j \in \mathcal{S}(\text{root})} |\hat{r}_j - \hat{r}_{\text{root}}|^p. \]  

(5)

**Optimization**

Optimization entails the search for solution(s), \( \hat{f} = \{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_n\} \), that minimize the objective function, \( W(t_1, t_2, \ldots, t_n) \). The minimization of this objective function when \( W \) is given by equations (1)–(5) is a nonlinear optimization problem which cannot be solved analytically. However, standard numerical techniques such as Powell’s method (Press et al. 1992; Nash and Stofer 1996) have proven to be efficient (requiring on the order of minutes of computation time on a midrange UNIX workstation) for phylogenies of up to 50 taxa. It is possible to explicitly calculate the derivatives of the objective function, which can be used in gradient-based optimization methods, but the analytical overhead required to do so does not seem warranted in view of the relative efficiency of Powell’s method in this problem.

As always with nonlinear optimization, considerable caution must be exercised in finding and testing local solutions for convergence. Moreover, the possibility of multiple solutions cannot be overlooked. To address these potential problems, all searches were started from different initial guesses at the parameters. Local stability of solutions was checked by perturbing them and restarting the search. If multiple optima were suspected, the local stability of each was checked and the size of the basin of attraction was estimated by perturbing them in a series of progressively larger perturbations until the algorithm converged to a different optimum. No real multiple optima were found in the simulations or analyses of real data sets.

**Including Fossil Constraints**

This optimization approach lends itself readily to including detailed information about fossil taxa. Without information from fossils, the optimization procedure described above can only provide estimates of dates that are relative to the unknown age of the root node in the tree. If that or any other node can be assigned an age, then absolute ages can be inferred.

However, in general, fossils do not fix the ages of internal nodes; they merely constrain them to be minimum ages (Norell 1992; Doyle and Donoghue 1993; Smith 1994). In particular, to obtain a minimum age for some node \( X \), it is necessary to have a fossil that belongs to a clade descended from \( X \). A fossil that appears similar to \( X \) or related to \( X \) does not suffice, because it could merely be a sister group (or even more distantly related) and therefore its age would not constrain \( X \) at all, because \( X \) is not ancestral to it. For example, the abundant fossil record of eudicot angiosperms, a large clade of angiosperms, puts an absolute minimum age on the root node of angiosperms themselves at about 125 MYA (Doyle and Donoghue 1993). If such information is available for internal nodes in a phylogeny, it should be incorporated into the optimization procedure described above. Fortunately, standard constrained optimization methods are reasonably straightforward to implement.

Constraints typically will consist of a condition such as

\[ t_i > \tau_i, \]  

(6)

where \( t_i \) refers to the time of internal node \( i \) and \( \tau_i \) is the age of the fossil that constrains it. The optimization problem can now be summarized as minimizing \( W \) subject to the constraints in equation (6). Some optimization terminology is useful here. A point, \( t = \{t_i\} \), is said to be “feasible” if it satisfies all of the constraints in equation (6). The possible solutions form a feasible “space,” and the problem is to optimize \( W \) within this feasible space of potential solutions. This is particularly difficult when the solution(s) lie on the boundary of this space.

Nonlinear constrained optimization methods are more difficult than any other sort of continuous optimization problem. However, so-called penalty function methods (Nash and Stofer 1996) seem to work well with the particular objective function described in equation (2). The basic strategy is to optimize a different function, \( \Phi \), related to \( W \). This new function is like \( W \) except near the boundaries of the feasible region, where it gets very large in a way that prevents the algorithm from leaving the feasible space and hence violating constraints. By sequentially changing the function \( \Phi \), any solutions on the boundary can be discovered in due course to within some tolerable error.

Define the sequence of functions \( \Phi^{(\theta)} \), such that

\[ \Phi^{(\theta)}(j) = W(t_1, \ldots, t_m) + \alpha^{(\theta)} \sum_r \frac{1}{t_r - \tau_r}, \]  

(7)

for \( r \) constraints on \( m \) internal nodes. Each node might have up to two constraints: a minimum age and a maximum age. There are also the implicit constraints imposed by the tree structure itself—that the age of a descendant node is less than or equal to the age of its immediate ancestor. The \( \alpha^{(\theta)} \) are a sequence of progressively smaller constants. The algorithm works by beginning in some part of the feasible space and using \( \Phi^{(1)} \) in a new optimization problem. Because \( \Phi^{(1)} \) gets infinitely large as the solution gets close to violating any constraint, the algorithm is guaranteed to stay within the feasible space. By repeatedly invoking this procedure with a sequence of smaller values of the constant \( \alpha^{(\theta)} \), the algorithm converges on the actual constrained solution as a sequence of approximate solutions, \( \{\hat{f}^{(1)}, \hat{f}^{(2)}, \ldots, \hat{f}^{(\theta)}\} \), where the \( \theta \)th solution is chosen by some appropriate fractional tolerance stopping criterion (Nash and Stofer 1996).
An intuitive sense of this can be obtained in the following way. Imagine a bowl-shaped surface (the optimization function) intersected by a vertical cylinder, which represents the constraints. The feasible solutions now lie on that part of the bowl within the cylinder. This is a surface with a boundary around it. Now, instead of trying to find the solution within that boundary itself, which is made difficult by what can sometimes be a very irregular boundary, we stretch the shape of the bowl so that instead of coming to an abrupt interface with the cylinder it increases smoothly upward, never quite getting to that boundary. The minimum on this new surface can be found by standard methods. By changing $\alpha$, we progressively make that smoothly stretched region match the boundary between the cylinder and the bowl more and more closely until, for all practical purposes, it looks just like the original case.

Error Estimation

A bootstrap resampling procedure can be used to assess error in the age estimates due to character sampling (Efron and Tibshirani 1993). For each of $N$ bootstrap replicates, the data matrix is resampled, branch lengths are estimated on the given tree (the same tree in all replicates), and then these are input to the algorithms described above. The distribution of the resulting age estimates for each node is examined and its standard deviation is calculated (table 1). The same procedure can be implemented with or without constraints.

Simulation Analysis of Accuracy of the Algorithm

The accuracy of the nonparametric smoothing algorithm was assayed by computer simulation. Trees of a given number of taxa were generated by a stochastic pure birth (Yule) process (Harris 1964; see Sanderson and Donoghue [1996] or Mooers and Heard [1997] for review of phylogenetic applications) and the time from the root to tips was scaled to unity. Durations of lineages in a Yule process are exponentially distributed (Harris 1964). Sequence evolution was simulated according to a nonhomogeneous Poisson process in which the rate of evolution was allowed to vary among branches but assumed to be the same across each of $L$ sites in a sequence. Any branch $k$ had a temporal duration, $T_k$, determined by the Yule process, and a local rate of evolution, $\lambda_k$, which was constant over the branch. The number of changes, or branch length, of branch $k$ was a Poisson random variable with mean of $L \lambda_k T_k$. A value of $\lambda = 1.0$ means that the expected number of substitutions at a site in a lineage from root to tip is 1.0. The rate itself evolved according to a random walk among $K$ rate “states” ($K = 11$ was sufficient to model a highly variable set of rates). Technically, the rate was assumed to evolve as a Markov chain with a fixed transition probability, $p$, and a state set of $(\lambda_{\text{min}}, \lambda_{\text{min}} + \delta, \lambda_{\text{min}} + 2\delta, \ldots, \lambda_{\text{max}} = \lambda_{\text{min}} + (K - 1)\delta)$. Thus, the rate in a branch was determined by the rate in its ancestor’s branch, and was either the same (with probability $1 - \rho$) or jumped up or down by an amount $\delta$, the rate increment (with probability $0.5p$). The possible rates were bounded above and below by $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$, and $\lambda_{\text{max}}$ was set to $(K - 1)\lambda_{\text{min}}$ in all runs. The rate of evolution was initialized at the root node at $0.5\lambda_{\text{max}}$. When $p = 0$, the rate is constant over the tree, and there is a perfect clock and a high rate autocorrelation in time. If $p = 1$, there is a minimal relationship between the rate in one branch and the rate in its descendant branches, and autocorrelation is low (although not zero, because the rates are still evolving along a Markov chain).

The case of perfect data, or infinitely long sequences, was simulated by allowing the rates to evolve as indicated and then using some large number for $L$ (generally 1,000) and setting the branch lengths to exactly $L \lambda_k T_k$ instead of a Poisson deviate with that mean. This is correct because the coefficient of variation in a Poisson process goes to zero as $L \to \infty$, which means the stochasticity can be removed.

A complete simulation consisted of $N = 250$ replicate Yule trees (15 taxa each) in which sequence evolution was allowed to proceed once on each tree according to a given set of the parameters $L$, $\rho$, and $\delta$. The tree and its corresponding branch lengths were input to the NPRS algorithm and the set of estimated divergence times, $\hat{t} = \{\hat{t}_i\}$, obtained. The accuracy or mean fractional error of the estimates was determined by

$$\frac{1}{m} \sum_{i} \frac{\left| \hat{t}_i - t_i \right|}{(\hat{t}_i + t_i)}.$$ (8)
where \( m \) is the number of internal node times and \( \{ t_i \} \) are the "known" divergence times read from the Yule trees.

For comparative purposes, an algorithm that assumes a molecular clock was subjected to the same battery of tests. A comparable clock-based method is that of Langley and Fitch (1974; see also Nei 1987), which takes estimated branch lengths and a Poisson-based model of character evolution and uses maximum likelihood to simultaneously estimate the rate of sequence evolution and divergence times, \( \hat{t} = \{ \hat{t}_i \} \).

The accuracy of the NPRS algorithm is shown for a series of simulations in figure 2 compared to the accuracy of the clock-based Langley-Fitch method. Accuracy was examined as a function of the rate transition probability \( \rho \), the magnitude of the rate increment \( \delta \), and the sequence length \( L \). Figure 3 indicates the portion of the parameter space in which the nonparametric smoothing method performs better than the clock-based method.

An Example from Seed Plant Evolution

As an illustration of the algorithms described above, divergence times within seed plants were estimated using sequence data from the chloroplast \( rbcL \) gene, which has been widely used for both phylogenetic analyses of plants (Chase et al. 1993) and reconstruction of ages of key events such as the origin of angiosperms (Martin et al. 1993; Goremykin, Hansmann, and Martin 1997). Details of this analysis will be reported elsewhere (unpublished data). Relative-rates tests and likelihood ratio tests reveal that \( rbcL \) is not evolving with rate constancy across land plants (see also Gaut et al. 1992).

The phylogeny is based on a synthesis of hypotheses of phylogenetic relationships in land plants. Branch lengths were estimated with maximum-parsimony methods (Swofford et al. 1996). Unlike in some previous work using this gene (Martin et al. 1993), all positions in codons were included in estimating branch lengths. The green alga, \( Chara \), was used as an outgroup to provide branch length estimates for the first two ingroup branches and was then removed from all subsequent trees. If this is not done, there is no way to partition branch lengths between the two branches descended from the root node of a tree. The tree and branch lengths are shown in figure 4.

The root of the tree was fixed in the Late Ordovician, at 450 MYA, which is consistent with the earliest paleobotanical evidence for land plants, which is the appearance of meiospores in tetrads (Graham 1993). In constrained optimization analyses, two additional fossil dates were used. One was the putative minimum age of a clade of seed plants comprising conifers and \( Ginkgo \) (labeled "CON" in figs. 5-8), which dates to the Carboniferous (320 MYA; Stewart and Rothwell 1993); the other was the minimum age of eudicot angiosperms (labeled "EUdic"), a subgroup of angiosperms ("AG") comprising most of the angiosperm diversity, whose distinctive tricolpate pollen appears at about 125 MYA at the Barremian-Aptian boundary of the Cretaceous (Doyle and Donoghue 1993).

Software Implementation and Availability

The unconstrained and constrained NPRS algorithms described in this paper are implemented in the author's program "r8s" available by anonymous ftp from loco.ucdavis.edu in directory pub. The program consists of ANSI C source code and can be compiled on UNIX machines with a C compiler (and probably on PCs and Macintosh computers, although this has not been tested). The program reads Nexus formatted files (Maddison, Swoford, and Maddison 1997) and also performs miscellaneous relative-rate tests. The example data described below are also included in a sample data file.

Results

Accuracy of Algorithms

The first column in figure 2 shows simulation results for data in which the sequence length is so long that the inferred number of changes along a branch is an exact estimate of the actual expected number of changes on that branch. A rate increment, \( \delta \), of 0% corresponds to no change in rate, or completely clocklike evolution (fig. 2, top graph in column 1). Both methods reconstruct divergence times without error in that ideal case. This means that both methods are statistically consistent under a clock; they converge to the true parameter values given enough data (Lindgren 1976). As the rate increment, \( \delta \), increases (indicating larger changes in rate over time), and the rate transition probability, \( \rho \), increases (meaning rate autocorrelation decreases), the accuracy in time estimates is higher for the NPRS method than for the clock-based method (fig. 2, middle graph in column 1). Eventually, as the rate increment and transition probabilities become too large, the clock-based method begins to outperform the NPRS method, presumably because of its robustness (fig. 2, bottom graph in column 1).

The second and third columns in figure 2 indicate the accuracy with finite amounts of data. Even when evolution is clocklike (either \( \delta \) or \( \rho \) equal to zero) both methods have a positive error, but the clock-based maximum-likelihood method has somewhat lower error (as it should, because likelihood methods are optimally efficient when the model is true). However, when a clock is not present, the performance of the NPRS method is often higher, with large but finite amounts of data (column 2), but begins to degrade as the amount of data declines (column 3).

A somewhat different perspective on relative performance is obtained from figure 3, which indicates whether the NPRS or the clock-based method performs better for a specific combination of \( L \), \( \rho \), and \( \delta \). A decline in performance with decreasing amounts of data is indicated by the trend from top to bottom in these three plots.

Divergence Times in Seed Plants

Results from the unconstrained clock-based analysis are reported in figure 5 and those from an unconstrained NPRS analysis are reported in figure 6. These
Fig. 2. — Results of computer simulations concerning the accuracy of the NPRS algorithm (shaded circles) for estimating divergence times versus that of the clock-based maximum-likelihood algorithm of Langley and Fitch (1974) (open squares). Accuracy is measured by percent error averaged over all internal node times (see eq. 8). Each point is the mean of 250 replicate simulations on random trees generated by a Yule process. Columns represent sets of simulations with the same sequence length, \( L \); rows represent sets with the same value of step size, \( \delta \). Step size is scaled relative to the maximum allowed rate of character evolution in the Markov rate model, \( \lambda_{\text{max}} \). Standard errors of the mean are less than 0.014 in all simulations in all of the graphs.

Analyses differ most notably in the age estimates of various groups of seed plants, and somewhat less dramatically in subgroups of angiosperms. The estimated age of angiosperms does not differ much between the two: both place it in the Jurassic. The clock-based results indicate that the ages of the Ginkgo-conifer group (CON) and the crown group of cycads (CYC) are clearly much younger than the fossil record indicates. The existence of fossils of all these lineages well before their estimated ages (Jurassic and Cretaceous, respectively) clearly falsifies the age estimates and, by proxy, the molecular clock itself. Fossils indicate that some of these taxa had split in the Carboniferous, 150 Myr before the estimated divergence times based on molecular data. Indeed, the clock-based estimate for the age of seed plants is Late Permian, well after seeds first appear in the fossil record.

The unconstrained NPRS analysis fares somewhat better, putting both seed plant groups (CYC and CON) back into the early Triassic at least, and putting seed plants as a whole back into the Early Permian. It is possible that the crown group cycads represented in this analysis actually did split in the Triassic, but the conifers certainly split from the Ginkgo line well before the time indicated here, despite this estimate being 50 Myr older than clock-based estimates. Presumably, these differences with the clock-based estimates are due to an increased rate toward the angiosperms and gnetales, suggested by the phylogram in figure 4.

Certain angiosperm subgroups with a good fossil record provide additional tests of the methods. For example, the ages of two eudicot clades, one consisting of *Nelumbo* and *Platanus* and the other consisting of *Fagus* and *Carya*, are both reconstructed as being much too recent in the clock analysis (fig. 5). These taxa are unquestionably present in the Cretaceous, long before the mid-Tertiary age estimates of figure 5. With the NPRS algorithm, the node subtending *Nelumbo* and *Platanus* is estimated to be much older, consistent with the fossil record, but the *Fagus-Carya* clade is still too recent.

All analyses were repeated enforcing two fossil constraints: minimum ages for conifer-*Ginkgo* (CON) and eudicots (EUDIC), as described above (figs. 7 and 8). Not surprisingly, the estimated age of seed plants becomes much more reasonable; after all, it cannot be younger than Carboniferous if one of its subclades has
Estimating Divergence Times

Rate Transition Probability, \( p \)

FIG. 3.—Relative performance of the NPRS and clock-based Langley-Fitch methods in computer simulations as described in the text (see also fig. 1). A shaded circle represents a parameter combination in which NPRS had lower error (higher accuracy) than Langley-Fitch; an open circle represents a combination in which NPRS had higher error. \( L = \) sequence length. Note that for infinite sequence lengths (top diagram), the estimated error rate is extremely low along both axes of the graph, and the relative performance of the two methods is therefore difficult to assess accurately, being subject to large stochastic variations.

Discussion
Utility of the Method

Bootstrap estimates of standard deviations in divergence times are on the order of 5–10 Myr for the Langley-Fitch clock method and slightly higher for NPRS (table 1). This is not surprising given the usual efficiency advantage of likelihood methods. The mean age estimated across the bootstrap replicates can be compared with the raw estimates to give an indication of bias in the estimators (Efron and Tibshirani 1993), from which it is clear that both NPRS and Langley-Fitch are slightly biased toward giving ages that are too young. However, in view of the magnitude of the error variance, this does not appear to be a significant problem.

FIG. 4.—Phylogenetic tree of 36 land plant rbcL sequences, with branch lengths proportional to estimated numbers of substitutions. The tree is based on work reported elsewhere (unpublished data) and is included here merely to illustrate the methods described in the paper.

a minimum age in the Carboniferous. These are also in accord with a late Pennsylvanian time estimated from other chloroplast and nuclear gene data (Savard et al. 1994). Both clock and NPRS analyses now estimate a somewhat older age for angiosperms of about Late Triassic to Early Jurassic, which is consistent with many recent estimates of angiosperm ages using molecular clocks (Wolfe et al. 1989; Laroche, Li, and Bousquet 1995; Goremykin, Hansmann, and Martin 1997), but not with that of Martin et al. (1993). However, once again the clock method reconstructs a much too recent Cretaceous age for the Cycas–Zamia split, and a Tertiary age for both the Nelumbo–Platanus split and the Fagus–Carya split. The NPRS method reconstructs a much older age for the cycads and for the Nelumbo–Platanus split but is also off regarding Fagus–Carya.

Marchantia
Lycopodium
Equisetum
Osmunda
Polypodium

Vascular Plants
Cycas
Zamia
Ginkgo
Picea
Podocarpus
Taxus

Seed Plants
Saururus
Drimys
Magnolia
Eupomatia
Chloranthus
Austrobaileyia
Nymphaea
Calycanthus
Hersea

Angiosperms (AG)
Pachysandra
Ranunculus
Pisum
Fagus
Carya
Enkianthus
Nicotiana

Nelumbo
Platanus
Spathiphylum
Serenoa
Lilium
Oryza

Eudicots (FLIC)

FIG. 4.—Phylogenetic tree of 36 land plant rbcL sequences, with branch lengths proportional to estimated numbers of substitutions. The tree is based on work reported elsewhere (unpublished data) and is included here merely to illustrate the methods described in the paper.

Bootstrap estimates of standard deviations in divergence times are on the order of 5–10 Myr for the Langley-Fitch clock method and slightly higher for NPRS (table 1). This is not surprising given the usual efficiency advantage of likelihood methods. The mean age estimated across the bootstrap replicates can be compared with the raw estimates to give an indication of bias in the estimators (Efron and Tibshirani 1993), from which it is clear that both NPRS and Langley-Fitch are slightly biased toward giving ages that are too young. However, in view of the magnitude of the error variance, this does not appear to be a significant problem.

Discussion
Utility of the Method

The simulation results indicate that it is possible to reconstruct divergence times more accurately using the NPRS method than using a clock-based method under certain conditions, including (1) enough data, (2) non-clocklike evolution, and (3) moderate to high levels of rate autocorrelation. Of these three factors, only the first is actually or potentially under the control of the investigator. The second, however, can at least be detected using the wide variety of rate tests available today (Li 1997), and therefore it will generally be possible to pre-
dict when the NPRS method should be considered a viable alternative to clock-based methods. The third factor, however, is something about which very little is known. It has been argued from first principles that rate autocorrelation ought to exist (Gillespie 1991), but despite considerable discussion of rate variation, very little is known about rate autocorrelation in time. The reason, of course, is that rate autocorrelation estimates require estimates of absolute rates, rather than just relative rates. Although modern phylogenetic methods allow branch lengths to be estimated, absolute rates and hence changes in absolute rates cannot be estimated without knowledge of absolute divergence times (Springer 1995 and references therein). Unfortunately, the point of the present paper is to improve upon the estimation of such divergence times under precisely those conditions in which rates might be expected to vary—hence, the dilemma.

Although these conclusions are somewhat negative, the following two points should be considered. First, the estimation of divergence times when rate constancy is absent is quite plainly a hard problem. The degrees of freedom can increase much faster than the amount of data unless certain assumptions are made about the evolutionary process. Thus, the fact that the simulations indicate that the method does not work well under all circumstances is hardly surprising. Second, there is little in the way of a baseline for comparison of the performance of the NPRS method. It is difficult to compare the relative accuracy of the present method to those currently available, such as linearized trees (Takezaki, Rzhetsky, and Nei 1993), because simulation studies like those used to evaluate the NPRS method above have not yet been performed on those methods. Indeed, it is odd how little work has been aimed at assessing the accuracy of divergence time estimation methods in

FIG. 5.—Phylogenetic tree with ages estimated according to the Langley-Fitch clock-based method. Geologic time scale is given in terms of millions of years and is based on the time scale of Palmer (1983). LP = land plants; CYC = Cycads; CON = Conifers plus Ginkgo; AG = Angiosperms; EUDIC = Eudicots. The age of land plants is set to 450 Myr. See table 1 for bootstrap estimates of error in ages.
FIG. 6.—Phylogenetic tree with ages estimated according to the NPRS method. See table 1 for bootstrap estimates of error in ages.

general. Important analytical results are known for the confidence intervals on clock-based estimates when the clock holds (Nei, Stephens, and Saitou 1985; Hillis, Mable, and Moritz 1996; Steel, Cooper, and Penny 1996), but there are few such results (possibly none) when rates are permitted to vary over time.

Generalization to Site-Based Models

In this paper, estimates of branch lengths are used to provide estimates of local rates. These estimates are obtained from the original sequence data by whatever means are chosen by the investigator, including parsimony, likelihood, or distance methods. It should be possible to incorporate that sequence data more directly into the NPRS method without the intermediate step of branch length estimation. Ultimately, that may permit the incorporation of more detailed information about the substitution process, including rate variation across sites, which is known to affect divergence time estimation (Palumbi 1989; Yang 1996).

The method of “penalized likelihood” (Green 1987) couples likelihood estimation with a nonparametric smoothing procedure for parameters that are difficult to model. Consider some set of sequences $S$ and some model of sequence evolution $M$, which is the same for all sites (Rodriguez et al. 1990). Let the unknown parameters, $\theta$, of the model consist of a set of $m$ node times and $n$ branch rates; $\theta = \{t_1, t_2, \ldots, t_m, r_1, r_2, \ldots, r_n\}$. The likelihood of the model with its parameters, $L(M, \theta \mid S)$, could be constructed using standard recursive methods (Felsenstein 1981). By itself, this would not allow estimation of the parameters, however, because the likelihood function would not permit simultaneous estimation of some rates and times. However, optimization of $L(M, \theta \mid S) + \lambda W(\mathbf{r})$, where $\mathbf{r}$ is the vector of unknown local rates in $\theta$, and $W$ is the NPRS
function described in equations (1)–(2), might impose sufficient constraints on the problem to permit solution. Here, \( W(r) \) is not a function of the branching times (as it is in eq. 3), which enter only into the likelihood term. The constant \( \lambda \) is a weighting coefficient that tunes the optimization problem to be more or less influenced by rate autocorrelation. This is an area of current work, but the computational problems are not trivial.

Another promising approach is via hidden Markov models (HMMs). Hidden Markov models have been used to study sequence analysis (Hughey and Krogh 1996) and sequence alignment (Mitchison and Durbin 1995) and to model rate variation across sites in a sequence (Felsenstein and Churchill 1996). An HMM is a model in which some of parameters to be estimated (here, rate) are themselves determined by a stochastic process of change. For example, rate might be modeled to be in one of two alternative states, “fast” and “slow,” with a given probability of changing between these states. The process of substitution is then modeled by first determining what rate state the process is found along a particular branch, and next determining the number of substitutions according to the chosen rate. This matches closely the kind of Markov process used to model autocorrelated rate evolution in the simulations described above. It might therefore be expected to yield accurate estimates of divergence times. Henze and Sanderson (unpublished data) have examined this kind of model in the context of data consisting of inferred branch lengths, a Poisson substitution process, and a fixed rate-transition matrix. The magnitudes of the fast and slow rates can be estimated by maximum-likelihood at the same time as the divergence times are estimated, but estimation of the rate transition probability is more problematic. It appears that, as with estimation problems in HMMs generally (Rabiner 1989), the shape of the
likelihood surface is complex and multiple optima are common. Evidently, these optima correspond to discrete alternative “reconstructions” of the ancestral rate states. Further work is needed.

Rate Autocorrelation and Phylogenetic Inference

J. Kim (personal communication) has pointed out that the likely existence of rate autocorrelation in evolutionary history raises the possibility that it should be included in model-based methods of phylogeny reconstruction. This would represent a significant departure from the historical trend in the development of phylogenetic tools away from sensitivity to rate variation. Ultrametric methods such as UPGMA were shown to be sensitive to deviations from clocklike evolution, and likelihood, parsimony, and distance methods such as neighbor-joining are considered superior for the very reason that they permit tree reconstruction to proceed despite rate variation (Swofford et al. 1996). The counterpoint to that, however, is that any of these methods might readily reconstruct a considerable amount of very rapid rate change without any penalty incurred in their optimization criteria (or, in the case of neighbor-joining, any effect in the distance renormalizations). It is possible that for some data sets a nearly equally optimal tree may entail a much higher level of rate autocorrelation than the optimal tree, and hence be more “likely” under the background assumption that autocorrelation is more likely than completely random rate change.

Fossils and Constrained Optimization

Constrained optimization methods were used in this paper in the context of the NPRS algorithm and the Langley-Fitch clock-based algorithm, but this approach to including fossils in rate analyses is much more general. It can be used with any method of determining...
divergence dates that estimates parameters via optimization, including but not limited to tree-wide likelihood methods such as those described by Felsenstein (1988), in which a clock model of sequence evolution is used to reconstruct ages of all internal nodes of a tree. Unfortunately, none of the software packages currently implementing such methods allow constrained optimization.

It has been noted many times that the fossil record has a great deal to say about molecular evolution (Springer 1995). One reality of the fossil record is that, except when preservation is exceptionally complete, fossils tend to reliably impose minimum ages, but not maximum ages, on nodes (Doyle and Donoghue 1993). Hence, it does not reliably fix nodes to a particular point in time (Norell 1992); rather, it constrains them to lie within some interval. Recent quantitative paleontological work has shown how these intervals can be estimated statistically based on the detailed preservation history of strata near the first or last appearance times (Marshall 1990; Marshall and Ward 1996). Very few studies that use molecular clocks to date divergence times take this simple fact into account, except to note that the reliance on minimum ages may bias results in a particular direction. Wray, Levinton, and Shapiro (1996, note 22), for example, pointed out that their estimates of Metazoan age may be considerably inflated by their reliance on minimum ages as proxies for actual ages. Springer (1995) has shown how this general practice can lead to contradictory inferences about even the existence of rate constancy. Constrained optimization methods may provide a powerful framework for incorporating the temporal uncertainties often associated with the use of fossils to date nodes in phylogenetic trees.

Acknowledgments

I am grateful to J. A. Doyle for extensive discussions of paleobotany and plant phylogeny, to J. Kim and J. Gillespie for discussions of the algorithm, and to two anonymous reviewers and the associate editor for very helpful suggestions.

LITERATURE CITED


STANLEY A. SAWYER, reviewing editor

Accepted August 26, 1997